Division of Population Health - Data Classification Policy v1.0

Approved at IG Committee Meeting 2024-07-22

Background

The University of Sheffield has a central Information Classification & Handling Scheme which dictates how different types of information are to be stored by members of the University. Non-research data within the Division of Population Health will be held in accordance with the University policy, whereas research data will be held in accordance with this policy document rather than the University policy.

Our policy is influenced by the 2019 Alan Turing Institute published paper called ‘Design choices for productive, secure, data-intensive research at scale in the cloud’, which has since become a standard benchmark that a large number of cloud based research environments are holding themselves to.

Scope

This policy applies to all members of the Division who are responsible for managing data.

This policy is focused on risk bearing data, which is a wide umbrella term, mostly focusing on Personal data and commercial-in-confidence data but may include other types of data such as commercially sensitive data. The Division refers to risk bearing data, defined in the Division of Population Health Information Governance Policy.

This policy will assign ‘sensitive data tiers’ to risk bearing data, ‘Tier 0’ data is out of scope as this describes non sensitive data.

Process

Triage Process

When a network folder creation request is sent into Population Health Data Security (DS) the tier of data expected to be held within the folder will be required information.

For a folder where this information is missing, due to the folder being created prior to the introduction of this data classification policy, the information will be gathered as part of the routine folders checks completed by Population Health-DS.

The tier that a project requires is determined in accordance with the severity diagram flowchart developed and maintained by the secure data services (SDS).

Tier Based Restrictions

The tier assigned to a project will differ depending on the severity of project data; meaning that different tiers will require different security controls.

Tier Explanations

The tier system is specifically tailored to research data and is therefore ideally suited to our needs, more so than the University of Sheffield classification system. The below section will outline what the Tiers in the environment are and what data is to be used in them. If during triage it is unclear what tier a project should be placed in, then the below section has more detail which should be referenced to help with the tiering process. A flowchart is available upon request to help with the classification process.

Tier 1

- Tier 1 environments are those used to handle pseudonymised data, commercial in confidence data which provides a minor competitive advantage, or datasets available upon request which have specific data security, data handling or access control requirements
- Tier 1 environments are also suitable for commercial in confidence data, intellectual property data, or legally / politically sensitive data where the data being lost, stolen or misused would have little or no impact on individuals or the university
- Tier 1 environments are not suitable for pseudonymised datasets where a member of your research team has access to the pseudonymisation ID, or if the dataset is rich enough to let you identify people within it using only publicly accessible information

This is equivalent to the ‘Internal Information’ and potentially ‘Restricted Information’ classification of the University of Sheffield Data Classification policy

The most sensitive type of data we would expect to see in a Tier 1 environment:

- Banded and/or pseudonymised direct identifiers (e.g Year of birth 2000 - 2009)
- Banded and/or pseudonymised indirect identifiers (e.g Place of Treatment replaced with a Pseudo ID)
- Commercial in confidence, private third party, intellectual property or legally / politically sensitive data which is of trivial reputational value

Note: All data in a Tier 1 Environment MUST be Pseudonymised or Anonymised:

Tier 2

- Tier 2 environments are those used to handle pseudonymised datasets where a member of your research team has access to the pseudonymisation ID, or if the dataset is rich enough to let you identify people within it using only publicly accessible information.
  Data may only be classed as Tier 2 if these risks are being appropriately managed by the research team (e.g the Pseudonymisation ID is held in a separate environment with very restricted access, analysis is only done on slices of the dataset and not the whole thing)

- Tier 2 environments may also be used to handle confidential data, as long as that data is not national security data. This includes data which is provided as commercial-in-confidence, or intellectual property, which does not meet the criteria for Tiers 1 (i.e where the data being lost, stolen or misused would have a minor or worse impact on individuals or the university)

This is equivalent to the ‘Restricted Information’ classification of the University of Sheffield Data Classification policy

The most sensitive type of data we would expect to see in a Tier 2 environment:

- Pseudonymisation ID that includes Initials

- Banded and/or pseudonymised direct identifiers (e.g Year of birth 2000 - 2009)

- Unredacted indirect identifiers

Note: All data in a Tier 2 Environment MUST be Pseudonymised or Anonymised:

Tier 3

- Tier 3 environments are those used to handle pseudonymised data where a member of your research team has access to the pseudonymisation ID, or if the dataset is rich enough to let you identify people within it using only publicly accessible information. Data should be classed as Tier 3 where these risks cannot be appropriately managed by the research team (e.g the research team have onboarded known individuals onto the project and need to link data to them, analysis has to be done on a dataset where the direct and/or indirect identifiers cannot be suitably banded and/or pseudonymised)

- Tier 3 environments may also be used to handle data which is provided as commercial-in-confidence, or intellectual property, which does not meet the criteria for Tier 1 or 2 (i.e where the data being lost, stolen or misused would have a moderate or worse impact on individuals or the university)

- This tier anticipates additional need to defend against compromises from attackers with bounded capabilities and resources, such as hacktivists, journalists, criminal individuals or competent individuals.

- Tier 3 environments are for identifiable datasets that do not pose a major risk to the subject's health, safety or security if it were to be disclosed.

This is equivalent to the ‘Highly Restricted Information’ classification of the University of Sheffield Data Classification policy

The most sensitive type of data we would expect to see in a Tier 3 environment:

- Unredacted Direct and Indirect identifiers alongside small and/or low impact datasets (e.g a data breach will not cause harm and is unlikely to cause distress in individuals)
- Banded and/or pseudonymised direct and Indirect identifiers where members in the dataset can be identified by individuals on the research team

Tier 4

- Tier 4 environments are for identifiable datasets wherein the data being lost, stolen or misused would have major or worse impact on individuals or the university
- Tier 4 environments may also be used to handle data which is provided as commercial-in-confidence, or intellectual property, which does not meet the criteria for Tier 1, 2 or 3 (i.e where the data being lost, stolen or misused would have a major or worse impact on individuals or the university.)
- Tier 4 environments are suited for data which may be subject to attack by sophisticated, well resourced and determined groups, such as criminal groups or state actors.

This tier corresponds to the government ‘SECRET’ classification.

The most sensitive type of data we would expect to see in a Tier 4 environment:

- Unredacted Direct and Indirect identifiers alongside Special Category Data
- Unredacted Direct and Indirect identifiers alongside rich and/ or high impact datasets

Mid Project Re-Classification

There may be times during a project where through the unintended creation of new data, or through heavy analysis; the tier of a project may need re-assessing. This could be due to datasets not being as strongly pseudonymised as was first thought, or certain analytical techniques implied making the data identifiable. For whatever reason, it may be necessary to change the tier that a project operates in. Any such changes can be made by contacting Population Health-DS alternatively IAOs will be asked to confirm the tier as part of project reviews conducted by Population Health-DS.

Appendix 1 Direct and Indirect identifiers and anonymisation

The Information Commissioner's Office (ICO) has published guidance on anonymisation[1]. There is also useful information regarding removing and transforming direct and indirect identifiers within Smith et al. (2015)[2] and Appendix 2 of the UKCRC guidance[3]

For the purpose of our process, definitions are as follows

Direct identifiers

Name, address, full postcode, NHS number (or similar identifier), email address (if it contains other information about them e.g. name, date of birth), biometric data, facial photograph or comparable image / video, audiotapes (voice recordings), names of relatives, dates related to an individual (e.g. date of birth)

Direct identifiers with lower disclosure risk

Telephone number, email address (if now identifiable data included), initials, vehicle identifiers, medical device identifiers, internet protocol addresses

Indirect identifiers; may be identifiable if in combination or unusual

Place of treatment (e.g. hospital, site, GP surgery), partial postcode, health professional responsible for care (e.g. principle investigator, GP name), sex, rare disease or treatment, date of treatment / diagnosis / hospital visit, place of birth, socioeconomic data (e.g. rare occupation, place of work, income, or education), household and family composition, anthropometry measures, multiple pregnancies, ethnicity, year of birth or age (age is potentially identifying if the recruitment period is short and is fully described), verbatim (free text) responses or transcripts, religion.

Consider also Small denominators (population size of <100) and Very small numerators (event counts of <3)

---

[1] Anonymisation: managing data protection risk code of practice (ico.org.uk)

[2] Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials. Tudur Smith C, Hopkins C, Sydes M, Woolfall K, Clarke M, Murray G, Williamson P. April 2015.

[3] data_sharing_sop_guide_v1.1_-1.pdf (ukcrc-ctu.org.uk)